DOCUMENT RESUME

ED 104 363 IR 001 755

AUTHOR	Tucker, Chris
TITLE	EUDISED Network Interchange Requirements and MARC Record Conversion.
INSTITUTION	Council for Cultural Cooperation, Strasbourg (France).
REPORT NO	DECS-Doc-75-2
PUB DATE	6 Jan 75
NOTE	39p.
EDRS PRICE	MF-$0.76 HC-$1.95 PLUS POSTAGE
DESCRIPTORS	Bibliographies; Catalogs; *Computer Programs; *Data Bases; Documentation; Educational Research; Information Centers; Information Networks; Information Processing; Information Storage; *International Programs
IDENTIFIERS	*EUDISED; European Documentation and Information System; File Conversion; MARC

ABSTRACT

A generalized file-conversion program suite for the EUDISED (European Documentation and Information System for Education) information network, suitable for running on a wide range of mainframe computers, is suggested and described. Time and expenditures required to implement the system are assessed. Two other possible methods of handling the data are examined and found less satisfactory. It is proposed that the softward be designed with a view to processing scientific and technical information in other data bases. (Author/SK)

ED104363

1R

COUNCIL OF EUROPE
CONSEIL DE L'EUROPE

Strasbourg, 6 January 1975

DECS/Doc (75) 2
Eng. only

COUNCIL FOR CULTURAL CO-OPERATION

COMMITTEE FOR EDUCATIONAL
DOCUMENTATION AND INFORMATION

EUDISED Network Interchange Requirements and MARC Record Conversion

by

Chris Tucker
Systems Consultant, London

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

38.058
04.6

001 755

C O N T E N T S

		PAGE
1.	INTRODUCTION	1
2.	THE NATURE OF THE EUDISED NETWORK	2
3.	DISTRIBUTION OF FUNCTIONS	6
4.	FILE CONVERSION IN THE NETWORK	11
5.	STRUCTURE OF THE CONVERSION PROGRAMMES	15
6.	SOFTWARE DEVELOPMENT COSTS	29
7.	CONCLUSION	36

1. INTRODUCTION

From its inception the planners of the EUDISED information network have tried to incorporate lessons learned by other similar networks in the way that the basic building blocks of the EUDISED structure have been determined. It has been rightly assumed for example that the adoption of international standards, where they exist in a suitable form, will help to reduce time and effort involved in the construction of the overall information processing system.

There is a hierachy of levels at which standards can be utilised, and in general these levels are independent, in so far as standardisation at one level does not demand standardisation at another. Whenever non-standard information is permitted a special translating interface must be designed and implemented, and this inevitably leads to an increase in costs, both in the system development phase, and later in the normal processing phase when the interface has come into operation. The independent informatic levels involved in the EUDISED mechanised network, at which data is processed, will include, from the "lowest" level upwards:

Information Level		Contents
Subfield	-	a bibliographic data item in its most elementary form, with an identifying code associated.
Field	-	one or more subfields linked together in meaningful association, with a field identifying code.
Record	-	one or more fields linked together in meaningful association, with a record identifying code. In the EUDISED context the associated fields all describe characteristics of one entity of educational interest.
Stream	-	one or more records linked together in meaningful association, with a stream identifying code. The association being simply that all records in the stream are being transmitted from one node to another in the network.

All EUDISED output programmes running at a processing node in the network, producing listings of one kind or another (eg catalogues) for users of the system, are ultimately concerned with manipulating subfields, but the raw data itself will have arrived in the first instance embedded in streams of records from many different locations. If all sources of streams of records adopt different sets of associative rules and identifiers then the translation processes necessary to move down through the levels to the sub-field level, where the processing programmes must work on standardised elements,

./.

are complex and wasteful to implement. The proper use of standards and well designed conversion software can reduce these incompatibilities to manageable proportions.

This paper is concerned only with a small area of the expected EUDISED standardised software approach covering the feasibility of generalised software for conversion between different versions of records conforming to the ISO 2709 standard. Such records are referred to here as MARC records, and it is assumed that the format details are well known to the reader.

Before discussing the necessary requirements of the software it is necessary to make some assumptions regarding the shape of the network in order to put the need for this software into perspective, and this is the purpose of sections 2 and 3.

2. THE NATURE OF THE EUDISED NETWORK

The details of the network philosophy to be adopted by EUDISED are still under discussion. Although it is fairly clear what types of processing nodes will be required the difficult decisions arise when an attempt is made to allocate the functions that must be carried out for successful operation of the network, to the different types of node. These are assumed here to be an international centre, a national centre for each member state and a few sub-national centres.

What basically are these functions? We must start by distinguishing between specialised functions concerning the educational aspects of the information and these other functions solely concerned with the processing and transmission of the information throughout the network in appropriate form which do not depend on the subject area covered by the information. There is no reason to suppose that organisations suited to carry out one type of task will necessarily be most effective in carrying out the other.

2.1 Specialised educational functions

i. These include: collecting the bibliographic data from primary and secondary sources of educational information, cataloguing, indexing and abstracting; coding the data unambiguously on worksheets.

ii. Devising output structures for education - orientated indexes, abstract journals and other lists. That is, deciding on sequence, record selection criteria and all other aspects relating to the appearance of the outputs so that they will be appropriate to the needs of the educational community.

./.

iii. Constructing search profiles for SDI-type requests or retrospective searches. (Even with the most user-orientated retrieval language some residual specialised assistance will inevitably be required).

iv. Creation and maintenance of the EUDISED multilingual thesaurus, and the control of other similar specific standards, eg coding structures.

v. Defining data validation rules and standards for bibliographic data elements used in the EUDISED data base.

vi. Administration of extensions or reductions in the network's coverage of educational material.

All these functions involve expertise in the specialised field of educational information and would always require close involvement of educationists at all stages.

2.2 Processing functions

These subject independent functions include:

i. Transcription of new and amended bibliographic data records and search profiles to machine-readable form, eg by key boarding to paper tape, magnetic tape, or OCR input documents.

ii. Conversion of machine readable record streams to a common internal format for computer processing.

iii. Validation by computer programme of the bibliographic data records and search profiles.

iv. Maintaining computer-based files of current and archival material.

v. Processing indexes and other lists by selecting records from the computer-based data files and organising them appropriately on machine readable files in internal format.

vi. Producing man-readable versions of these indexes and lists in whatever form required eg microform, computer typeset, output, line printer listings, or on-line through a video terminal.

vii. Maintaining a computer based file of SDI profiles.

viii. Processing search profiles, SDI or retrospective, against current or archival material and producing outputs in the desired physical form.

./.

ix. Conversion of selected files in internal format to MARC format for transmission to other nodes in the network.

x. Producing the necessary accounting and billing information for all services provided by the network.

xi. Producing statistical and other management information concerning system operation, currency of material, demand on services etc.

xii. Document reproduction service for non-copyright material on microfile or paper.

All these functions are subject-independent and those that involve computer processing should be carried out by programmes which load the appropriate controlling tables on detecting signals in the data stream. Even validation, which comes the nearest to being subject dependent, can be handled more effectively, and generally, by specifying a set of validation routines appropriate to bibliographic data items, preferably under ISO supervision. The process referred to above in 2.1 (v) would then consist of allocating a standard validation routine number to each possible type of field to be input. The validation programme would then execute the appropriate standard, subject independent, routine, for each.

In the EURONET environment this type of procedure, when applied in all the processing areas mentioned above, would enable a software system to be devised that is almost completely subject transparent.

2.3 Transmission functions

This area is concerned with ensuring that the right data is made available at the right time at the appropriate processing point, where one or more of the functions outlined in 2.2 will be carried out. This can involve simple movement of data between two processing points within one node for example source data sheets, carried by hand between preparation room and key-boarding room in a national centre, or more complex inter-nodal communications.

At the present time discussions on the exchange of machine readable data, directly or indirectly, between national centres have assumed that the data would be carried on magnetic tape and transmitted by post, and indeed this would of course be a possible method. However in the time scale of development of the EUDISED network we can expect that much faster, and more convenient methods of transmission will be available, and although sub-national nodes can be expected to send magnetic tapes to their national centre, the national and international centres are much more likely to transmit data (bibliographic records and search profiles) through a packet switching network using the public telephone network in conjunction with appropriate private data transmission links.

./.

In the EURONET context all national nodes for educational information processing would be included in the data transmission network. Some nations would require separate nodes for processing in different subject areas, others would use one centre for concentrating the processing functions over many subject areas including education, but would require separate centres to co-ordinate the subject orientated aspects outlined in 2.1.

The writer has taken this likely development into account in the description of the conversion software where it affects the argument. In general however it has no effect because it really should not matter how the stream of data has been transmitted, as far as the processing programmes are concerned. MARC data arriving on magnetic tape through the post could be, say, spooled into a disc file using a standard manufacturers utility programme; alternatively MARC data arriving through a data transmission network could be directed into an identical disc file using another standard utility - all the extra "clothing" added to the record stream by the source node, for transmission purposes would be stripped away automatically by the communications software at the receiving node and would not reach the disc file. Hence the processing programmes accessing the file subsequently would not know, or need to know, by what means the data had arrived.

3. DISTRIBUTION OF FUNCTIONS

The allocation of the functions listed in section 2, between the various types of node in the network, will ultimately be decided by a combination of logical and political considerations. It has been agreed in the draft EUDISED programme for 1976-1978 that an international centre is an essential focus for the educational network and there is clearly a need for national centres to supervise the collection of source material and profiles from their geographical region, and to ensure that educational information is disseminated efficiently in that region. The need for sub-national centres will vary from country to country and are of course a matter for individual national consideration.

In a simplistic way then we can attempt to show how the functions listed above might sensibly be distributed.

3.1 The International Centre

3.1.1 Education orientated functions

Advising on suitable output formats, as referred to in section 2.1 (iii);

Control of EUDISED thesaurus and other coding standards - 2.1 (iv); Quality control of data - 2.1 (v); Control of coverage - 2.1 (vi).

3.1.2 Processing functions

Keyboarding occasional special items/corrections and profiles - 2.2 (i);

Conversion of data streams, received from national centres and international organisations, to internal format - 2.2 (ii);

Validation of input local to the international centre - 2.2 (iii);

Maintaining reference files of all material including material received but not disseminated, because it is of limited local interest in one national area only, eg non-book materials which can only be lent within one national boundary, perhaps - 2.2 (iv);

Producing selective lists of international interest covering the whole range of material - 2.2 (v), (vi);

Carrying out occasional retrospective searches covering the whole range of material when this is not held by the national centre, or no national centre is appropriate - 2.2 (viii);

./.

File conversion from internal format to MARC format for transmission to national centres - 2.2 (ix);

Accounting (and billing) for the above functions - 2.2 (x);

Management information for the international centre - 2.2 (xi);

Limited document reproduction service - 2.2 (xii).

3.1.3 Transmission functions

Co-ordinating the receipt of data streams from the national centres and other sources, and the despatching of merged data streams back to the national centres.

3.1.4 Note that because of the logical separation of these three different types of function it may be sensible to have two international centres in the network. One organisation responsible for the education-orientated aspects, and the other for the processing and transmission functions; this latter centre is only likely to be economic if utilised by other subject areas, and it would use identical programmes for all disciplines.

3.2 The National Centres

One of these will be designated by each EUDISED member state and will carry out these functions:-

3.2.1 Education orientated functions

Responsible for ensuring that all relevant data originating within the national boundary is collected and encoded - 2.1 (i);

Deciding the precise output format required for lists and indexes to be published by the centre, taking into account the limits of the software available, and the advice of the international centre, taking into account language considerations - 2.1 (ii);

Constructing search requests for educationists within the national boundary which are subsequently processed by the national centre, or passed onto the international centre - 2.1 (iii);

Allocating responsibility for coverage of source material to sub-national centres (if any) - 2.1 (vii).

3.2.2 Processing functions

"Keyboarding" new national material and amendments onto a machine readable medium. (Bibliographic data and profiles) - 2.2 (i);

Conversion of MARC records from international and sub-national centres (if any) - 2.2 (ii);

Validation of bibliographic data and profiles - 2.2 (iii);

./.

Maintaining a local file of international material deriving from the international centre according to agreed selection procedures, plus, perhaps, material which the international centre has not been sent, because it is in a category judged by the international centre to be of insufficient interest - 2.2 (iv);

Producing indexes and lists of material held in the national files of particular national interest - 2.2 (v); (vi);

Maintaining a file of SDI profiles - 2.2 (vii);

Carrying out searches - 2.2 (viii);

Selective tape service for sub-national nodes - 2.2 (ix);

Accounting and billing for services provided by the national centre - 2.2 (x);

Management information for the national centre - 2.2 (xii).

3.2.3 Transmission functions

Organisation of the efficient movement of data between searchers of primary/secondary source material, encoder, keypunch operator, computer room, proof reading section etc, and the despatch of machine-readable new and amended material (including search profiles) to the international centre by post or data link;

Dissemination of search results, indexes etc to the individual or sub-national centre.

3.3 The sub-national centre

These centres will sometimes be appropriate to assist the national centre in the gathering and dissemination of information; for example where the necessary subject specialists are distributed geographically distant from the national centre in existing organisations.

(The ERIC network has 16 such clearing houses). Individual terminal users with access to the national data base for submission of search queries can also be regarded as sub-national centres for SDI purposes.

3.3.1 Education orientated functions

Collection and encoding of source material within the restraints defined by the national centre - 2.1 (i);

Submitting search requests to the national centre - 2.1 (iii).

./.

3.3.2 Processing functions

Certain large sub-national centres may transcribe worksheets into machine readable form for their national centre and direct entry terminals might be convenient instruments for this process. Search profiles could be entered in the same way. Magnetic or paper tape could also of course be used as the output medium for this process - 2.2 (i):

In other circumstances work sheets and profiles in manuscript form would be submitted (by normal postal services or special delivery) for processing at the national centre.

It is unlikely that sub-national centres with their own in-house computer processing power will be necessary simply on cost-effective considerations, but terminal facilities will be almost unavoidable for efficient operation of the network.

3.3.3 Transmission functions

Limited in most situations to the movement of data between searcher of the primary/secondary sources of information and encoder, and finally to the national centre for processing.

Dissemination of search results to the individual user.

3.4 Summary

It is clear then that the main burden and diversity of transmission functions will be undertaken, not surprisingly, by the national centres, and the subject-dependent functions will be shared in the main between international and national centres, but these types of function do not concern us in this paper.

The processing functions which are subject independent are however of immediate interest. Of the 12 areas identified in section 2.2, 10 involve suites of computer programmes, and of these 10 areas the international centre carries out 9 of the functions and each national centre all 10. Now if we assume that in the EUDISED network alone there will be say 15 national centres participating then how inefficient and expensive it will be to allow the haphazard development of 15 variants of suites of programmes when one well-planned generalised version of each suite would be of universal applicability.

When this argument is extended to the EURONET context where many more centres in many other subject areas might also be involved in separate programme development the potential for wasted effort is raised by another order of magnitude. The 10 suites of generalised programmes, if carefully planned, should be able to handle all the subject areas without difficulty.

./.

The special needs of individual national centres, or subject areas, are catered for in the formulation of individual tables, thesauri etc which drive the common software in a particular way. The interfaces with different processing main frame computers can be dealt with in a similar way, once only for each variant species.

./.

4. CONVERSION IN THE NETWORK

This and subsequent sections are concerned only with a limited processing area; those functions referred to in 2.2 (ii) and 2.2 (ix), covered by the terms of reference for this paper.

4.1 In the description of processing functions in a foregoing section a distinction was made between internal and external file formats. This is because there is no "a priori" reason why the MARC format should be specially suited to the internal processing requirements. In fact, though in practice many organisations have used the format, usually in a slightly amended form, for their main bibliographic data base this has normally, one suspects, been a decision by default. It so happens that where main files are held in a serial, or sequential form, in a conventional record by record manner, a record structure that includes a directory offering direct access to each data field, is an essential requirement for efficient data processing and it is often convenient to adapt an existing format with this characteristic rather than designing a special one from first principles.

However, the days of simple serial main files in bibliographic data processing are disappearing fast. The bibliographic data bases of the future are likely to be organised in an altogether different, more complex way, designed for efficient processing of on-line queries, and therefore requiring some degree of direct access capability, as well as for normal batch processing needs. For example: (a) certain data elements may be held (only) in inverted, or transposed files for fast retrieval processing, and (b) it will probably be necessary to hold local variations of some record fields and this will involve special links between certain fields in the structure.

Apart from these file structural considerations there are control elements in the MARC format that are not necessary in every record in the data base, though they may need to be held as part of some kind of file header; data packing techniques will also permit internal space savings to be made.

With a generalised software system a standard internal file and record structure will need to be defined and adopted by the EUDISED network.

Individual standard implementations on different main frame computers would follow, and each would interface with the common software through file handling routines that would be reasonably machine independent. We have assumed of course that the generalised software will be written in a common high level language, with the assistance of certain assembler level routines written once for each main frame, to improve efficiency in areas like character string processing.

./.

Figure 1 shows a possible flow pattern for new and amended records in the EUDISED network and indicates the types of conversion software that will need to be available.

4.2 The international processing centre we can assume would receive new material for the data base in three format categories: EUDISED Standard MARC, MARC (but not EUDISED Standard), non-MARC.

4.2.1 Non-MARC formats can only be conveniently handled by special purpose programmes written explicitly one for each source of material. It is to be hoped that in the long term, with wider acceptance of MARC, such inconvenient sources will disappear. It is suggested here that such non-MARC files are converted to EUDISED MARC format by special programmes (B), each written explicitly for one non-standard file and record layout.

4.2.2 Several sources can be expected to submit data in their own MARC implementation that does not conform to EUDISED standards. For example certain on-going international organisations and some national centres, cannot be expected to change their current file structures just to satisfy the EUDISED ideal requirements. It will be a EUDISED responsibility to convert such files to the EUDISED standard format, and this can be accomplished by using a standard MARC to MARC programme, along the lines described in section 5 (ie Programme A).

4.2.3 Where national centres have been set up after the agreement on EUDISED standards, with systems based on EUDISED software, then a particular implementation of the exchange format will be standard. A standard EUDISED MARC to Internal Format programme can be used to handle all such files, and the outputs from programmes A and B, and can convert thereby all incoming data to a common internal form (Programme C).

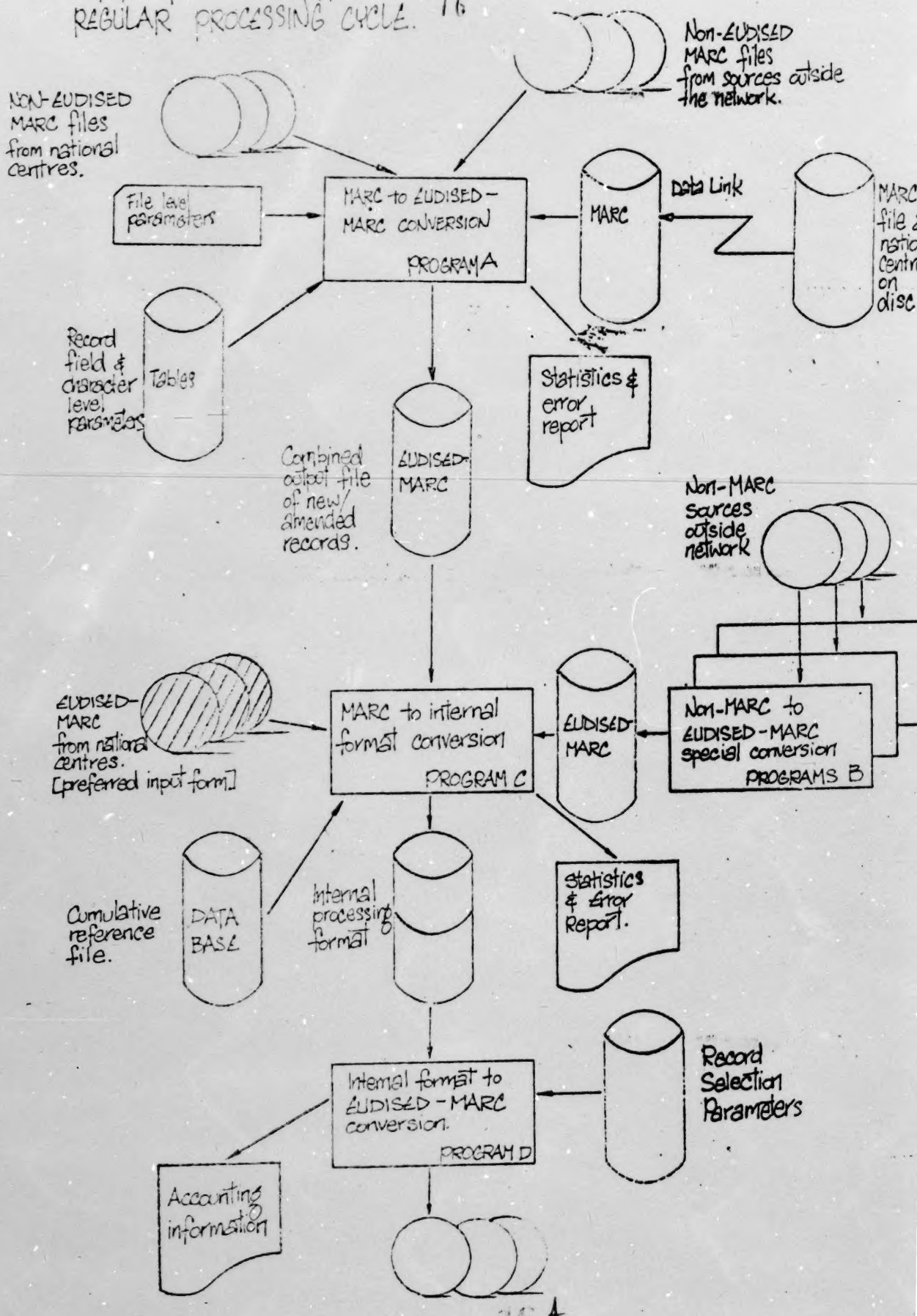
4.2.4 When Programme C has merged all new material and the data base has been successfully updated (by a later programme), selection of material for exchange purposes can be carried out by Programme D. Normally a copy of all new material and amendments/deletions will be sent to each contributing centre, plus other subscribers.

The data record structure on all outgoing files will be EUDISED standard MARC and Programme D does the necessary internal format-EUDISED MARC conversion.

4.2.5 Note that a national centre using EUDISED software can also use Programme C to convert the incoming exchange file of new material to internal format for subsequent processing. Other recipients will need to write their own special programmes.

./.

INTERNATIONAL CENTRE.
REGULAR PROCESSING CYCLE. 16


Non-EUDISED MARC files from sources outside the network.
NON-EUDISED MARC files from national centres.
File level parameters
MARC to EUDISED-MARC CONVERSION PROGRAM A
MARC
Data Link
MARC file
Record field & character level parameters
Tables
Statistics & error report
Combined output file of new/amended records.
EUDISED-MARC
Non-MARC sources outside network
EUDISED-MARC from national centres. [preferred input form]
MARC to internal format conversion PROGRAM C
EUDISED-MARC
Non-MARC to EUDISED-MARC special conversion PROGRAMS B
Cumulative reference file.
DATA BASE
Internal processing format
Statistics & Error Report.
Internal format to EUDISED-MARC conversion. PROGRAM D
Record Selection Parameters
Accounting information

4.3 The programmes referred to above would be table-driven and in the EURONET context would be used, without change, for all scientific and technical information systems in the network. Programme C, is in a sense a redundant step in a purely EUDISED environment because one might design Programmes A and B to convert directly to internal processing format. But in taking a wider view and knowing that each subject area will demand its own implementation of MARC, it is more convenient to arrange the processing steps in the way indicated. For it will be possible to parametrise the output format of Programme A, and the expected input format for Programme C to suit the subject area being processed without fundamental changes to the software logic.

5. STRUCTURE OF THE CONVERSION PROGRAMMES

Only Programme A is described in any detail here. Programmes B, C, D are not within the terms of reference of the present study.

5.1 Programme A. MARC to EUDISED-MARC Conversion

This programme converts a data stream containing records conforming to ISO 2709 format standards to the EUDISED implementation of ISO 2709.

ISO 2709 is concerned with the logical structure of a record only and not with data content; the necessary format details about the particular implementation are described in the fixed length part of a record itself and this enables a programme written for a MARC record in general, to locate required fields or sub-fields in a specific implementation. The fields which convey this information are indicator length, identifier length and directory map; the fact that ISO 2709 specifies that these fields shall be present in every record implies that on a standard MARC file they may change from record to record within one data stream. In practice this just does not happen and it is difficult to conceive of a situation where it would be necessary.

Perhaps it is a weakness in the standard that such redundant information should fill several character positions in each record. (A better solution might be to carry a standard "header" record, as the first record in any data stream, which would give general information concerning the stream as a whole eg date written/ transmitted, maximum logical record size, character set, blocking factors, list of tags used, and the format description fields mentioned above. Such a header record could be used by any processing programme to set up internal parameters in the programme for control of subsequent processing of the rest of the file of records).

In any case a conversion programme would be too cumbersome if it has to allow for such variations in the record structure within a single data stream and so for our purposes we assume that these four fields are constant for the stream and that the programme will obtain this information from the first MARC record in the stream.

The conversion programme works then by using this stored directory map to locate the start address of each in-put data field in turn, via the directory, and to convert this field, sub-field by sub-field, to an appropriate character string for placement under an appropriate field and sub-field code in the converted output record.

To effect the conversion a number of tables will need to be constructed at the beginning of the programme and held in care for the duration of processing. If the programme is to handle the conversion of a number of different files consecutively

./.

then those tables referring to input files will need to be loaded at the beginning of each new file input sequence.

5.1.1 These parameter tables are required at the levels of file, record, field, sub-field and character.

File Level

A series of parameters will need to be read at programme run-time to define the sequence in which the different MARC input files will be presented to the programme. This is an area which, being at the most machine dependent level, is most difficult to generalise. However these run-time parameters on cards will be divided between job-control language statements, and programme data in a way that depends on the machine on which the programme is executing. The information to be conveyed for each file includes:

a. Source identifier, eg National centre code number.

b. The storage medium on which the file is normally held, eg disc or magnetic tape, and file name.

c. Data recording information ie 7 or 9 track tape; phase encoded or NRZI, etc.

d. Character set in use on file, eg EBCDIC, ISO 7- bit.

e. Header labels, present or not.

f. Number of reels (when a multireel file).

g. Blocking information, eg maximum block size, long record handling method.

This information is input in a coded form, and it enables the programme to set up the correct switches and table loading parameters to process the file in the correct way.

At end of file the next set of parameters are read and the initialising process repeated. It is assumed that the output file will always be in a standard (EUDISED) form so the necessary parameters to describe it can be preset in the programme. However in the EURONET context, using the same programme, certain output file parameters would be read at the beginning of the run.

Record Level

This information would be held on backing store and would be loaded using the source identifier code to locate the appropriate set of conversion rules, stored previously.

./.

Required information includes:

a. Conversion rules on how implementation codes from the input records should be translated for output. The input field is four characters long and it should be sufficient to indicate for each character position the following options:

- Copy "as is"

- Replace by a specified character, if a specified character value, or range of values is present. (Repeat for each range of values).

- Always replace by a specified character.

b. Directory map for output records, ie how many characters in a directory entry are used to hold "length of data field", and how many hold "starting character position".

Field Level

This information would not change from week to week so it would reside on backing store and be loaded on detection of source identifier code. It would include:

a. Conversion rules on how to allocate field output numbers (tag numbers) to each field that may appear in the input record directory. For each possible input tag value the options are:

- Copy specified field to an output field of specified tag number. (This may be the same number as at input).

- Delete specified field.

A default option is inadvisable here so all possible input tag values must be specified in the table. A special procedure for handling fields 001 and 002 in a standard way must be devised.

b. Conversion rules on how to translate indicator values. Indicators are usually used to give general information about the field eg its provenance, level). This is a list of possible input values with the equivalent required output values for each field.

A default option can be specified giving a standard output value for all input indicator values not specified explicitly. The number of indicator positions will be fixed by EUDISED for the output file.

./.

Sub-field Level

These parameters are associated with their corresponding field number (tag value) and must describe the conversion rules on how to allocate output values to sub-field identifiers appearing in the input record. Such identifiers, if present, are introduced by the FS1 of ISO 646. This table therefore consists of a list of possible identifier strings for each field with their output equivalents. In some circumstances it may be necessary to move certain sub-fields within one input field to different output fields eg $a of 123 to go to $c of 456 but $b of 123 to go to $q of 789. This facility must be allowed for in the tables.

The basic options are, as at the field level:

- Copy sub-field with specified identifier to specified field and give it the specified identifier.

- Delete sub-field with specified identifier.

Character Level

In an ideal world all EUDISED input MARC files would be generated by the source in the preferred EUDISED internal character set so that character conversions would be unnecessary. Although this can perhaps be ensured as far as EUDISED national centres are concerned it will almost certainly be inevitable that character conversion facilities will be essential for files from outside the network.

The conversion is most conveniently carried out in three stages.

Stage 1 - On reading the record a character by character conversion of the whole record, from "record length" field to "record separator", is carried out. The aim of the conversion is to make the data concerned with record structure intelligible to later routines in the programme. This means that characters are translated by direct table look up to a EUDISED standard bid pattern so that numerics and separators at least are recognisable. The actual data fields, after this translation, may still be in a non-standard format because the pattern of control codes may be different in the input file from that in use by EUDISED.

This intermediate EUDISED form resulting from the first translation will need to be based on an 8-bit character pattern, to cope with 6, 7, 8-bit incoming codes from 7 or 9 "track" files, on a "one-for-one" basis.

./.

The "7-bit" codes mentioned will be on files from EUDISED national centres, if the current EUDISED proposal to adopt ISO 7 - bit encoding on exchange files is accepted. In fact there seem to be no advantages in the choice of this system, other than that it is an international standard; it falls between the two stools of 6 and 8-bit systems, satisfying neither capability efficiently. A better choice might be an 8-bit coding system, possibly based on EBCDIC, which could also become an international standard character coding system, (preferable perhaps to the 8-bit extension to the 7-bit standard). This paper is not the place to argue this case; however one can say that if an 8-bit intercharge system was adopted, and national processing centres were based on byte machines then the software development costs would be significantly reduced, not only on the programme under discussion, but also throughout the software system.

A character conversion table of this simple kind will need to be set up, once only, for each variant of input character set presented to the programme. The tables would be stored on direct access store and loaded when required.

Stage 2 - The character (Stage 1) translation means that directory and fields can be accessed by standard programme routines; however for copying across a sub-field from input to output record, a facility must be provided to ensure that bibliographic data fields contain character strings that are in a standard EUDISED form for subsequent processing in the system as a whole.

Because of the large size of the character set to be represented in EUDISED man-readable outputs, and the complexity of some output requirements, control codes will eventually have to be introduced into the character string in a standard manner to describe in a logical way such features as non-roman alphabets, diacriticals, print formatting character, character sizes and mathematical formulae. It is unlikely that all non-EUDISED sources will adopt the same principles for control characters, so a special translation sub-routine will need to be written for each variant to convert the input character strings to a EUDISED output string.

The software should be written in such a way that there is a software interface at which characters heading for the output string can be intercepted and processed by the final conversion routine.

Stage 3 - This routine enables a character (string) in EUDISED standard format to be replaced by a different string at output; this is a change not of format, but of data content.

./.

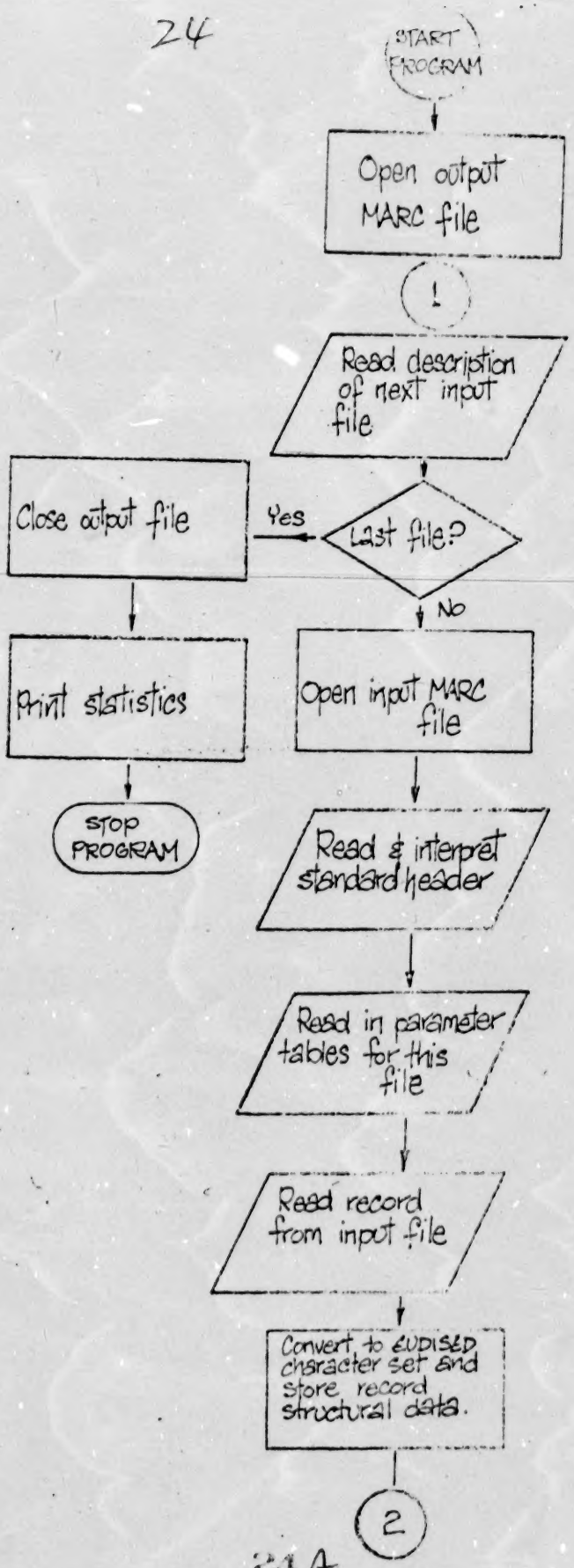
For example we might wish to change the internal representation of greek "α" to "ALPHA" in the output string because of some processing limitation of the recipient computer, or output printing system, compared to the better input capability of the source; another similar likely requirement is the deletion of certain control codes. This editing could be carried out automatically by a standard routine, driven by a conversion table drawn up for each input file type.

It is unlikely that this facility would be used at all by the international centre, using Programme A to deal with the input of new data, because its job should be to gather the records at the highest level of information content that the sources can provide, which would imply that no degrading of the quality (the function of stage 3) would be in the interests of the centre.

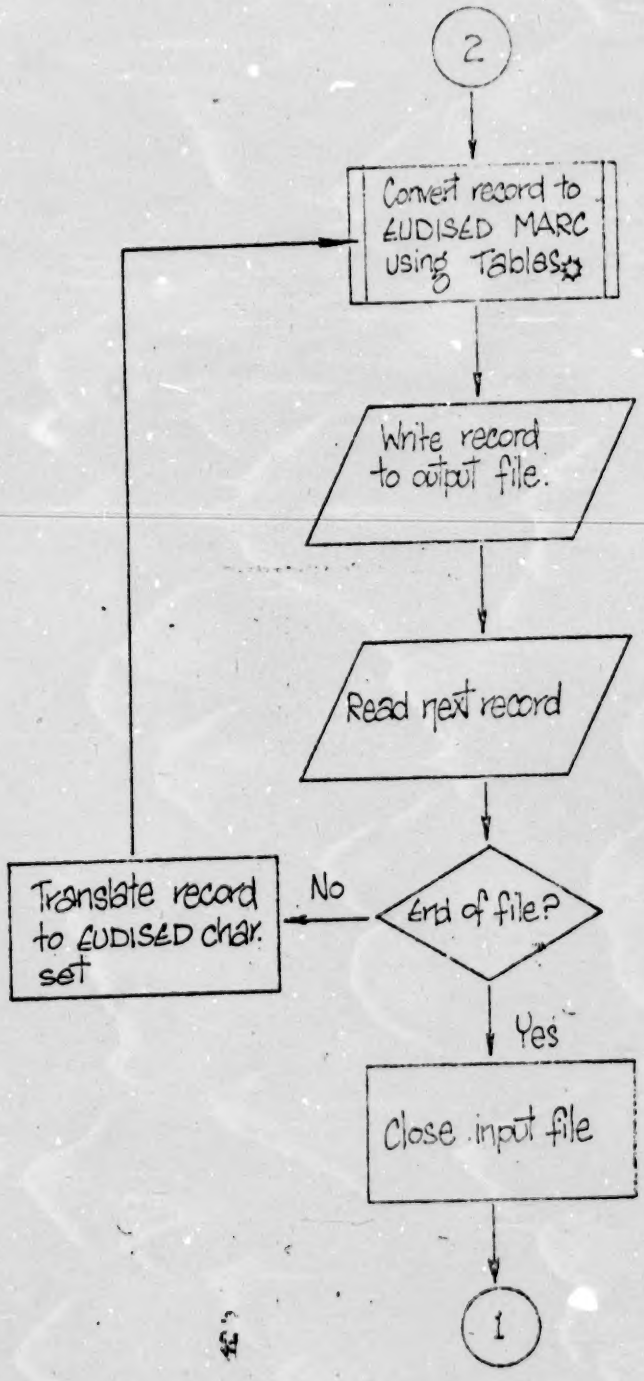
However when, for example Programme D is used in the reverse direction to provide a "combined" tape containing "new" material, for the national centres it may be EUDISED policy to edit the data to some extent, using the stage 3 facility to cater for something nearer to the lowest common denominator of the national centres' requirements. Alternatively the national centre receiving the EUDISED tape may use the stage 3 conversion facility to reduce the complexities on the file to something that its own processing programmes can handle.

./.

24


START PROGRAM
Open output MARC file
1
Read description of next input file
Last file?
Yes
Close output file
No
Print statistics
STOP PROGRAM
Open input MARC file
Read & interpret standard header
Read in parameter tables for this file
Read record from input file
Convert to EUDISED character set and store record structural data.
2

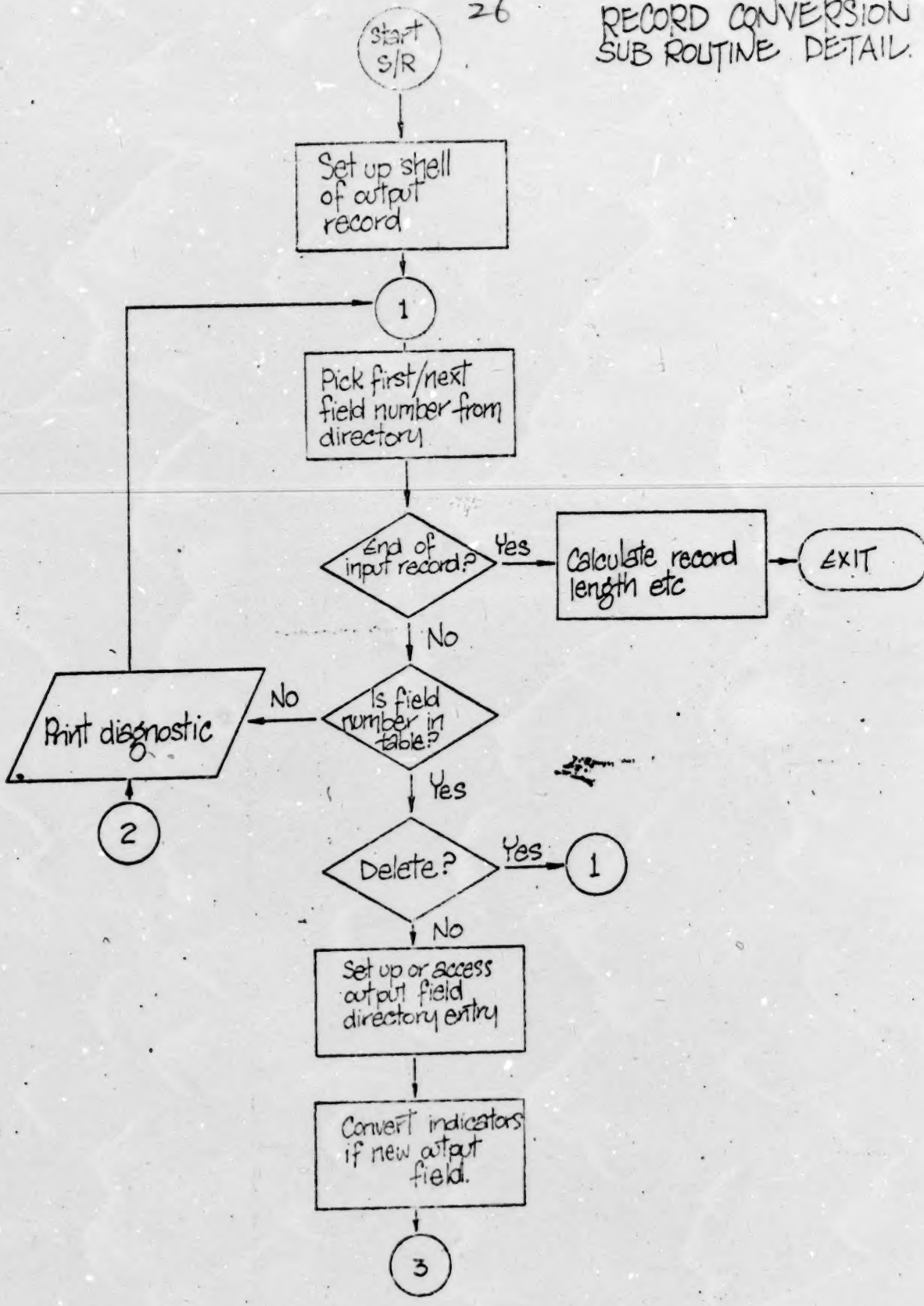
24A

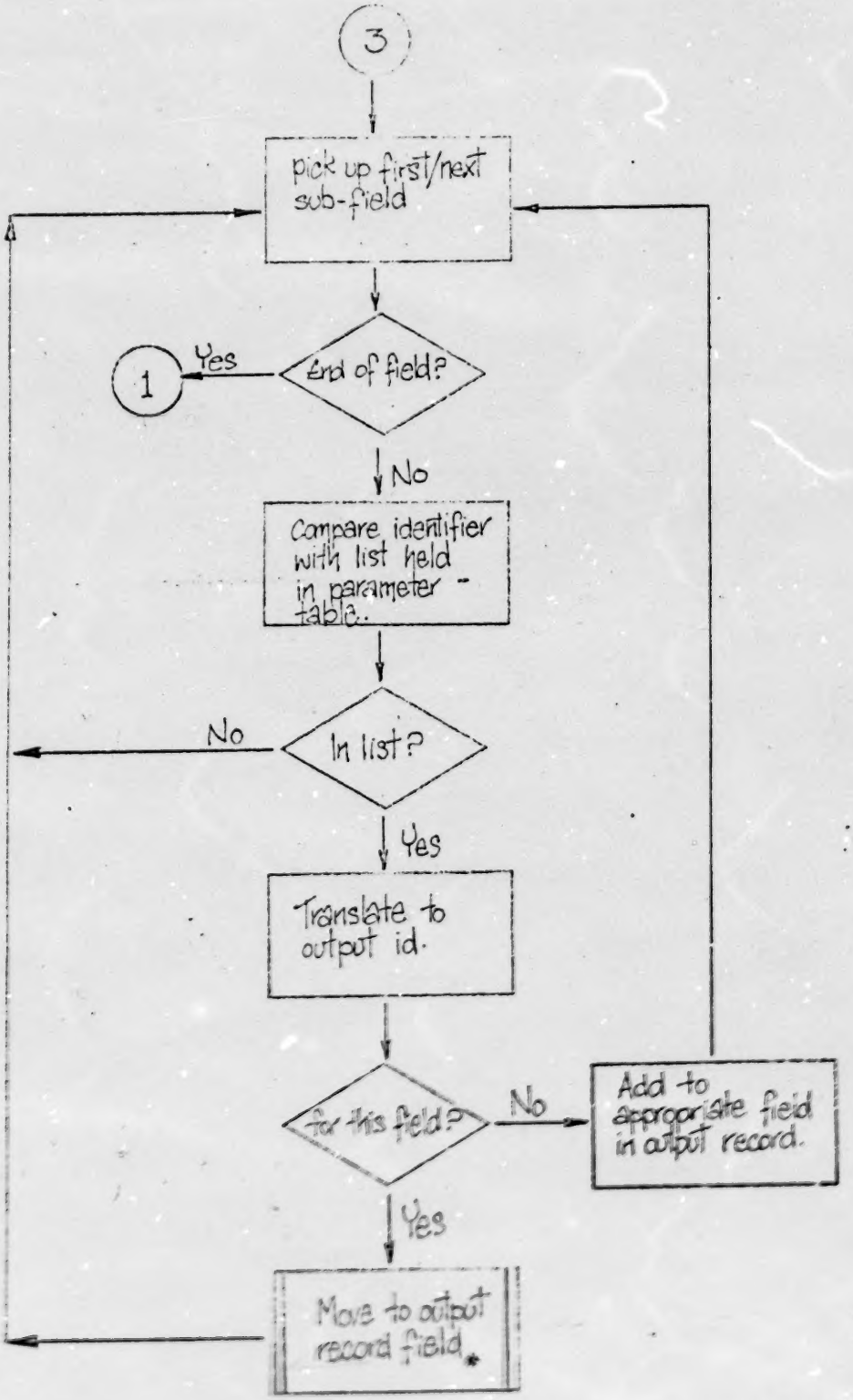


☼ See detail flow chart on next page:

26

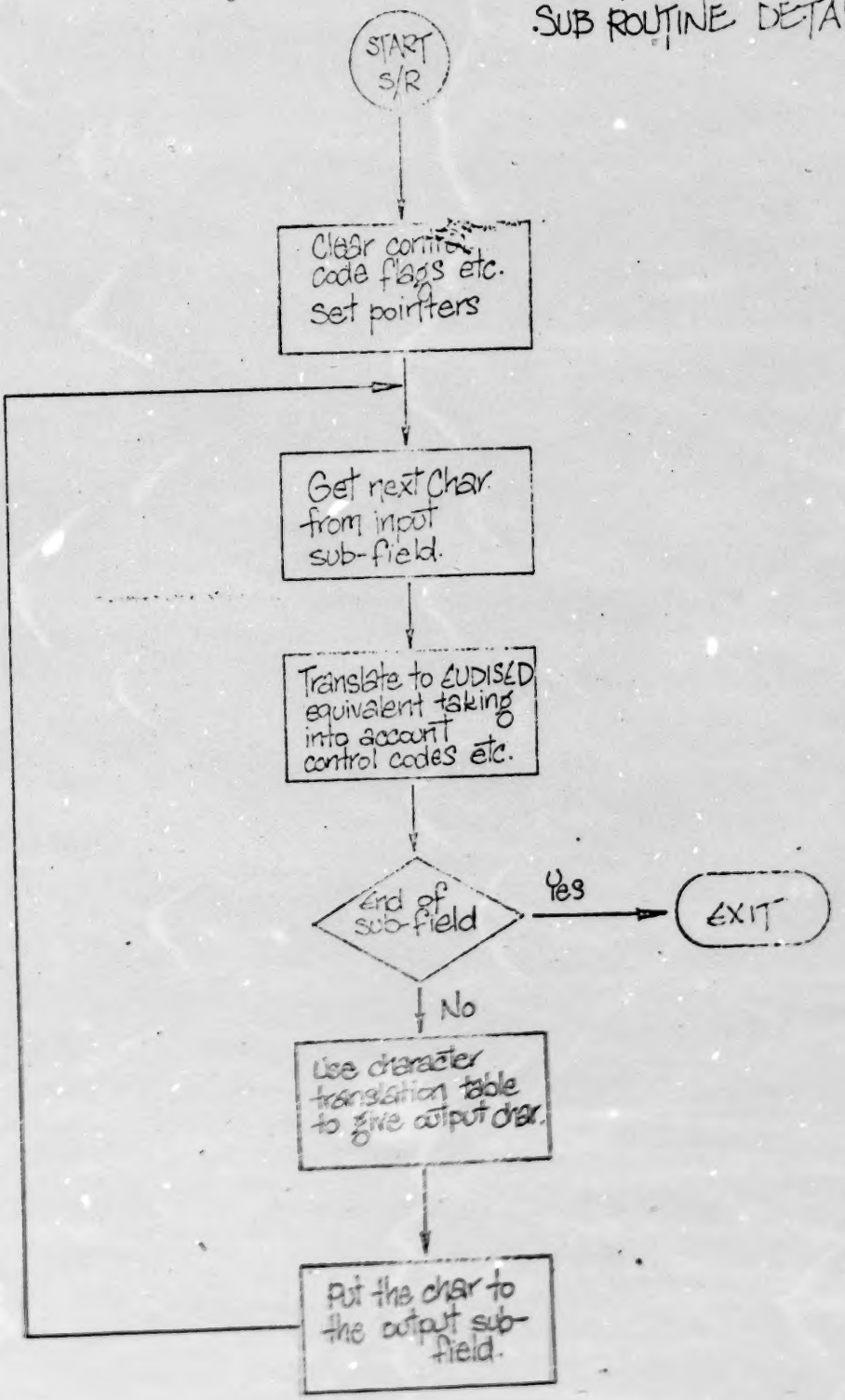
RECORD CONVERSION SUB ROUTINE DETAIL.





*See detail on next page.

SUB-FIELD CONVERSI
SUB ROUTINE DETAIL.


START S/R
Clear control code flags etc. set pointers
Get next Char. from input sub-field.
Translate to EUDISED equivalent taking into account control codes etc.
End of sub-field
Yes
EXIT
No
Use character translation table to give output char.
Put the char to the output sub-field.

5.2 Programmes Type B Special Conversions

Unless it is found to be possible to persuade all sources of records for EUDISED to conform to the MARC standard there will be a definite requirement in the output system for special purpose programmes, to be written by EUDISED to convert each non-standard data stream to EUDISED-MARC format. The complexity of this task will depend entirely on the format of the input records and the degree of correspondence between EUDISED MARC fields and those of the non-standard source.

A recent (18 November 1974) letter from the ERIC Processing and Reference Facility, in connection with this paper, indicates that at the present time the "ERIC format" to MARC format programme previously in use, has now been discountinued due to "inefficiencies"; this was a special conversation programme, in the sense used here. In a production EUDISED system if ERIC is one of the external sources chosen, then this existing conversion programme could probably be modified at little expense and run by ERIC, at source, or by the international centre on receipt of the ERIC format tape.

In other situations a conversion programme will need to be designed from scratch for each non-standard file. The output from each programme would be a standard EUDISED-MARC file with appropriate descriptive file labels or header record, identical in form to the output from Programme A.

5.3 Programme C. EUDISED-MARC to Internal Format Conversion

This programme as figure 1 shows, reads standard EUDISED-MARC files and converts them to a convenient form for internal processing. It carries out certain authority/validation checks by accessing the bibliographic data base, allocates accession numbers to new material, and locates existing accession numbers for the deleted and amended records when necessary.

Field validation checks will normally have been carried out by programmes at the national centres and ideally validated fields could be flagged in one of their indicator positions, to identify the programme routine used but in certain circumstances, eg for particular "low grade" sources, additional validation will need to be carried out by the international centre.

This programme will probably also carry out some kind of "dummy" up-date of the data base to ensure that the amendment/deletion information is consistent with the current state of the data base.

Output from the programme will be a combined file of material from all sources, MARC and non-MARC, in an efficient record structure, suitable for rapid processing in the system. Statistics on numbers of records for each source, currency of data etc will be printed out along with error details. (Errors can be located and corrections resubmitted in an amendment cycle not described here.)

Note that each national centre will also use this same programme, by passing one or two irrelevant facilities, to convert the regular combined exchange tapes received from the international centre, as a preliminary to the processing that maintains their own data bases.

5.4 Programme D. Internal Format to EUDISED-MARC Conversion

This programme would be used by the international centre to select material for transmission to national centres in exchange format. The same programme would be used both for the regular weekly or fortnightly selection of current material and also for occasional requests for copies of large sections of the data base; for example when a new national centre comes into being and requires the whole data base, or specific categories of material only, within given accession number range.

The programme will therefore have a front-end which carries out the selection of records on the basis of parameters supplied for each national centre. The parameter types would include:-

a. Accession number range to be covered by the selection;

b. Categories of material to be considered eg books but not film strips;

c. Availability eg only select if the item may be ordered for examination outside its country of origin.

After records have been selected they are processed according to criteria laid down by the international centre and probably not subject to national centre option. That is the records are standard EUDISED MARC, with a specific character set and field number allocation. It is quite possible, for example, that the international centre may not pass on certain fields which may be considered of too local an interest to be included on the exchange file; or perhaps if one national centre, technologically ahead of the others, provides data with a very large character set, this may be downgraded before distribution to other centres unable to process it, though it would be held in the full form by the international centre.

./.

This programme would also gather statistical and accounting information.

- Note that this conversion programme would also be used by most national centres to generate regular exchange files containing new material for transmission to the international centre. (These centres would be using the same internal record format, and generalised EUDISED software for the input/validate/update cycle for processing new material and Programme D would select from the national centre's new material those records required by the international centre for dissemination.)

./.

6. SOFTWARE DEVELOPMENT COSTS

6.1 Programming Lanugage

In designing software for use in an international context, operating on a number of different computer types, it is clear that programme portability must be one main consideration, otherwise it will be necessary to write the same programmes as many times as there are different computers. In the development of the conversion software a high level language should therefore be used whenever possible. This may be supplemented by character string handling routines written in the appropriate assemblers to improve efficiency in certain areas. However, as far as EUDISED is concerned, because the quantity of records being converted is relatively small, the actual cost benefit of more efficient processing by assembler code subroutines may be out-weighed by the extra effort required to write multiple versions of these routines in the first place. It may be more expedient to write all the programmes in the chosen high level language, only introducing assembler routines if and only if the chosen language is quite incapable of performing some specific function.

Which language should be used? We must choose one for which a compiler is available on all the computers to be used by the national centres, this probably eliminates PL/1, and leaves the decision to be made between Algol, Cobol and Fortran. One suspects there is little to recommend one rather than another, none are ideally designed for the kind of character string manipulations and character modified table look-up facilities required in the software: a detailed examination of the language facilities required, at programme specification time, would enable the most suitable to be selected. On the present evidence ANSI COBOL seems a good choice, for the self-documenting machine independent language. Let us assume that this language has been chosen as the EUDISED standard, and let us consider in that context the conversion programmes described above.

Programme A. MARC to EUDISED-MARC. This is used at one location only, the international centre, so it is not essential that it be written in COBOL. However it could use COBOL programme procedures common to other programmes so is probably best written in that language.

Programmes B. These are used only by the international centre and should be written in assembler language.

Programme C. EUDISED-MARC to Internal. This is used by most participating centres so would be written in COBOL.

Programme D. Internal to EUDISED-MARC. Used by most centres so should be written in COBOL.

./.

Use of a high level language will of course mean that programme development and maintenance costs can be considerably reduced from what would be incurred in developing multiple assembler versions of the programmes.

6.2 Use of International Standards

In the estimate that follows it is assumed that exchange records input to the MARC to MARC conversion programme conform to the following standards:

a. ISO 2709. Documentation - Format for bibliographic information interchange on magnetic tape. (Currently under revision).

b. ISO 1001. Magnetic tape labelling and file structure for information exchange.

It would be convenient to extend this standard to include other general information about the bibliographic information on the file in optional fields left available for such purposes by the standard formulators, eg character set in use on the file.

c. ISO 646. 7-bit coded character set for information processing interchange. (International reference version).

d. Block spanning rules are under consideration by ISO/TC97 and any standard arising from these discussions should be incorporated in EUDISED standards.

6.3 Basis of Estimates

6.3.1 The writer was unable to find any helpful published references to similar generalised file conversion programmes which might have assisted in estimating time and costs. The figures that follow are based on the writer's experience, in particular with reference to non-generalised "near MARC" to MARC software designed and implemented by the author for the exchange tape service offered by an established scientific technical information dissemination centre. In addition some discussion of the British Library Software Package with its originators gave another basis for comparison.

6.3.2 Estimates are given in terms of man weeks, and assume that staff involved in the project are of "average" ability.

6.3.3 Estimates given here assume that the software development is purely for EUDISED purposes; it has been argued earlier that a generalised EURONET orientated package that can handle EUDISED files would be a sensible approach; the cost premium to pay in these circumstances would be perhaps only 30% in addition to the figures quoted.

./.

6.3.4 The estimates do not include very variable costs associated with the development of special conversion programmes for non-MARC files (Programmes B).

6.4 Estimates of Development Effort

Stage 1 - Systems Analysis: investigating the precise requirements of the conversion software, establishing file structures, table formats and print layouts. Defining the interface requirements for successful integration of the conversion suite with the EUDISED internal processing software.

Senior Analyst	-	4 weeks
Analyst/Programmer	-	3 weeks
		7 weeks

Stage 2 - Programme Specifications: preparation of detailed programme descriptions.

Common assembler routines		-	2 weeks
Programme A	MARC to EUDISED-MARC	-	3 weeks
Programme C	EUDISED-MARC to internal format	-	2 weeks
Programme D	Internal format to EUDISED-MARC, with selection facility	-	3 weeks
			10 weeks

Senior Analyst	-	5 weeks
Analyst/Programmer	-	5 weeks

Stage 3 - Programme Writing: coding the assembler language routines for the first host machine

	-	2 weeks
Coding Programme A in a high level language		3 weeks
" Programme C " " " " "		3 weeks
" Programme D " " " " "	-	3 weeks
		11 weeks

Analyst/Programmer	-	6 weeks
Programmer	-	5 weeks

Stage 4 - Programme Testing: setting up programme testing data, both parameter tables and files, and testing out the programmes.

Common routines	-	3 weeks
Programme A	-	7 weeks
Programme C	-	5 weeks
Programme D	-	7 weeks
		22 weeks
Analyst/Programmer	-	11 weeks
Programmer	-	11 weeks

Stage 5 - System Testing: establishing that the programmes are consistent in combination as a complete conversion system by testing thoroughly with large data files

	-	4 weeks
Analyst/Programmer	-	2 weeks
Programmer	-	2 weeks

Stage 6 - Documentation:

Preparation of systems and programme documentation and user manual

	-	4 weeks
Analyst/Programmer	-	2 weeks
Programmer	-	2 weeks

In addition, supervision/project management would require perhaps 1 day or week for the duration of the programming stages of the project

	-	5 weeks
Grand total of		63 weeks effort

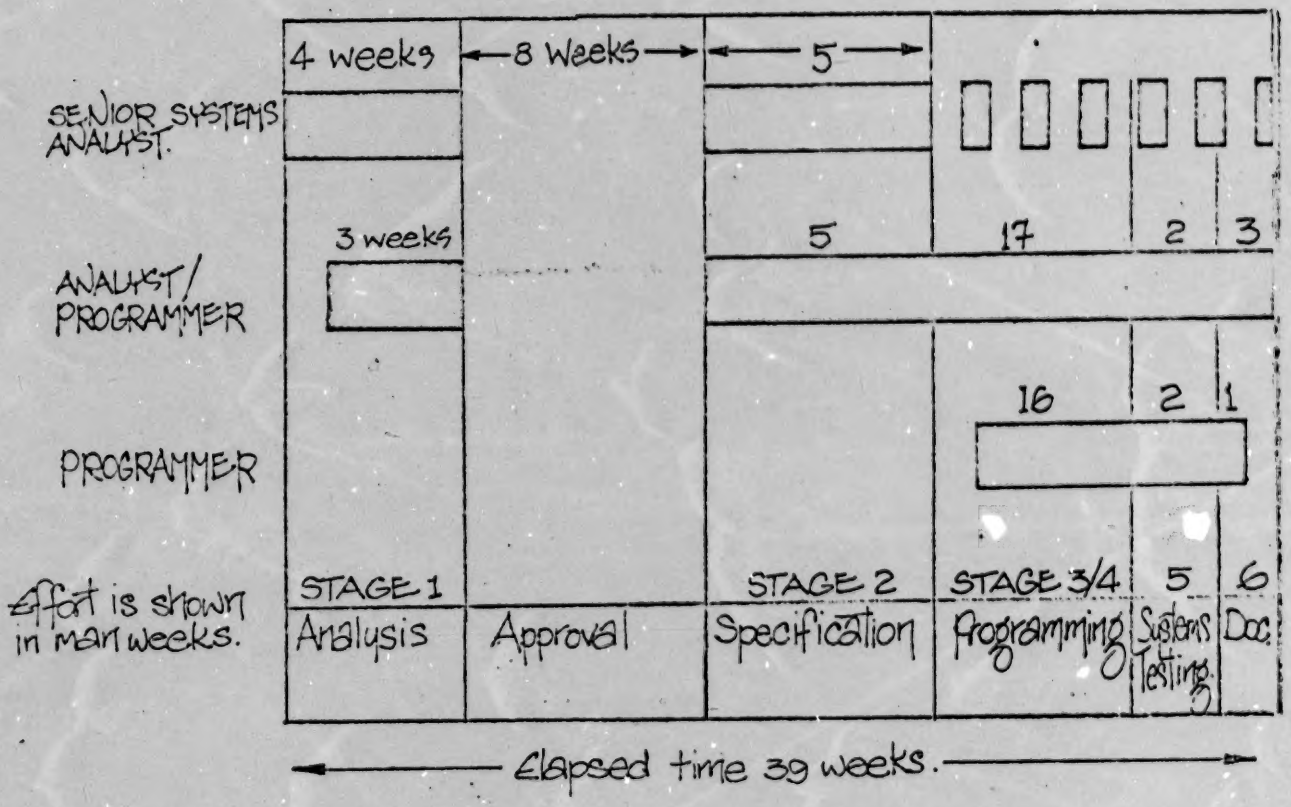
./.

Implementation of the software on second and subsequent different computer main frames would require perhaps 16 man weeks of additional programming effort per new version.

6.5 Implementation Schedule

The conversion software described above could reasonably be implemented in an elapsed time of 10 months as shown in the diagram on p. 58. This includes a break of 2 months after the detailed systems analysis has been carried out to allow adequate time for discussion and acceptance of the proposals, by the member states.

./.


4 weeks
8 Weeks
5
SENIOR SYSTEMS ANALYST.
3 weeks
5
17
2
3
ANALYST/ PROGRAMMER
16
2
1
PROGRAMMER
Effort is shown in man weeks.
STAGE 1
STAGE 2
STAGE 3/4
5
6
Analysis
Approval
Specification
Programming
Systems Testing.
Doc.
Elapsed time 39 weeks.

In the schedule above it is assumed that the senior analyst will continue as project-manager when systems work is completed, and that an analyst/programmer will be involved full-time throughout the project.

The total effort involved is as follows:

Senior Analyst	-	14 weeks
Analyst/Programmer	-	29 weeks
Programmer	-	20 weeks
		63 weeks

At current rates for software consultancy services the total expenditure on staff costs can be expected to be in the range of £15,000 to £17,000.

6.6 Other costs

The two other areas of expenditure are data preparation and computer processing for programme tests. The former is not significantly expensive, and is unlikely to exceed £250. Computer processing costs are difficult to estimate without knowing more details of the programmes involved, and the first host machine, but are likely to be in the region of £1,000 to £1,500.

./.

7. CONCLUSION

Study of the software system requirements indicates that the initial implementation of a generalised file conversion programme suite for EUDISED, suitable for running on a wide range of main frame computers, is technically feasible and could be completed in approximately 10 months elapsed time for an expenditure of some £15,000, involving some 60 man weeks of effort. Implementation of this software on different computer hardware would be a straightforward exercise.

The main alternatives to the generalised conversion software approach described are:

a. The imposition of a rigid definitive MARC standard on all suppliers of machine readable data files to EUDISED, whether national centres, international organisations, or private sources of educational information. This would eliminate the need for the MARC to EUDISED-MARC conversion and the "special" conversion programmes described. This possibility is judged to be unrealistic, and in any case would save at most one quarter of the conversion software cost, or

b. the treatment of each input exchange data file as an independent entity and therefore preparing a special conversion programme for each, on an ad hoc basis, to carry out the necessary conversion to internal processing format. This would involve as many programmes as there are sources of educational information in machine readable form, but would eliminate the need for a MARC to EUDISED-MARC conversion, or a EUDISED-MARC to internal format conversion. The development cost would depend on the number of special programmes to be written. Optimistically the first conversion might take about 10 weeks to implement, with the second and subsequent programmes taking less effort as experience was gained. On the basis that this method would save perhaps one half of the development effort described above for the generalised system (no Programme A or C) eg 30 weeks, the break-even point would appear to be around four sources - above this number the generalised method is cheaper. Of course a system based on these ad hoc programmes would be inconvenient to operate and inflexible to changes in input file specification because format details would tend to be embedded in procedural statements rather than being held in tables stored as data to the programmes.

Neither of these alternatives appears to be preferable to the generalised approach discussed in this paper, particularly if the software is designed with a view to handling the processing of scientific and technical information, beyond the field of education, in a EURONET environment.